Evaluating Lustre Network Performance over IB and RoCE

Matthew Vandeberg (matvandeberg@gmail.com); David Medin (david@davidmedin.com); Benjamin Schlueter (ben.sch1@protonmail.ch) |
Supercomputer Institute team Cable Guys; HPC-DO | Mentors: Jesse Martinez; Dominic Manno; Doug Egan; Trevor Bautista; Devon Bautista

Introduction

TERMS:

RDMA: *Remote Direct Memory Access*
RoCE: *RDMA over Converged Ethernet*
IB: *InfiniBand*
IPoIB : *Internet Protocol over InfiniBand*

BACKGROUND

With the increasing performance of Ethernet, the possibility of replacing InfiniBand with RoCE in a Lustre network has become more feasible. Currently, high-performance computing relies on highly parallel network file access to maximize computational performance. To accomplish this, a Lustre file system is often used with an InfiniBand network to provide the interconnect speed required by many HPC applications. This project aims to evaluate the implementation difficulties and performance differences of replacing a traditional InfiniBand Lustre network with Ethernet that takes advantage of RDMA using RoCE.

GOALS

Topology		Baseline TCP/IP		Baseline RDMA		Lustre Network		Lustre File system	
		IPv4	IPv6	IPv4	IPv6	IPv4	IPv6	IPv4	IPv6
IB	unrouted	✔	✔	✔	✔	✔	N/A	✘	✘
	routed	✘	✘	N/A	N/A	✔	N/A	✘	✘
Eth	unrouted	✔	✔	✔	✔	✔	N/A	✘	✘
	routed	✘	✘	N/A	N/A	✔	N/A	✘	✘
IB+Eth		✔	✔	N/A	N/A	✔	N/A	✘	✘

✔: completed ✘: not completed

Cluster Specifications

HARDWARE

- **CPU:** AMD EPYC 7502 32-Core
- **RAM:** 128 GB
- **NIC:** Nvidia ConnectX-6
- **Switches:**
  - Nvidia MQM8790-HS2F
  - Arista DCS-7280PR3
- **Cables:**
  - Nvidia HDR (200 Gb/s) optical cables
  - ENET HDR (200 Gb/s) breakout cables

SOFTWARE

- **OS:** Rocky Linux 8.8 (kernel 4.18.0)
- **Lustre:** version 2.15.3
- **MOFED:** version 5.8-2.0.3
- **Open MPI**: 4.1.5
- **Benchmarking Software:**
  - `iperf` (version 2.1.6)
  - `perftest` (version 6.16)
  - `IMB` (Intel MPI Benchmarks)
  - `GPCNeT` (version 1.3)
  - `LST` (Lustre Network Selftest)

Network Topology

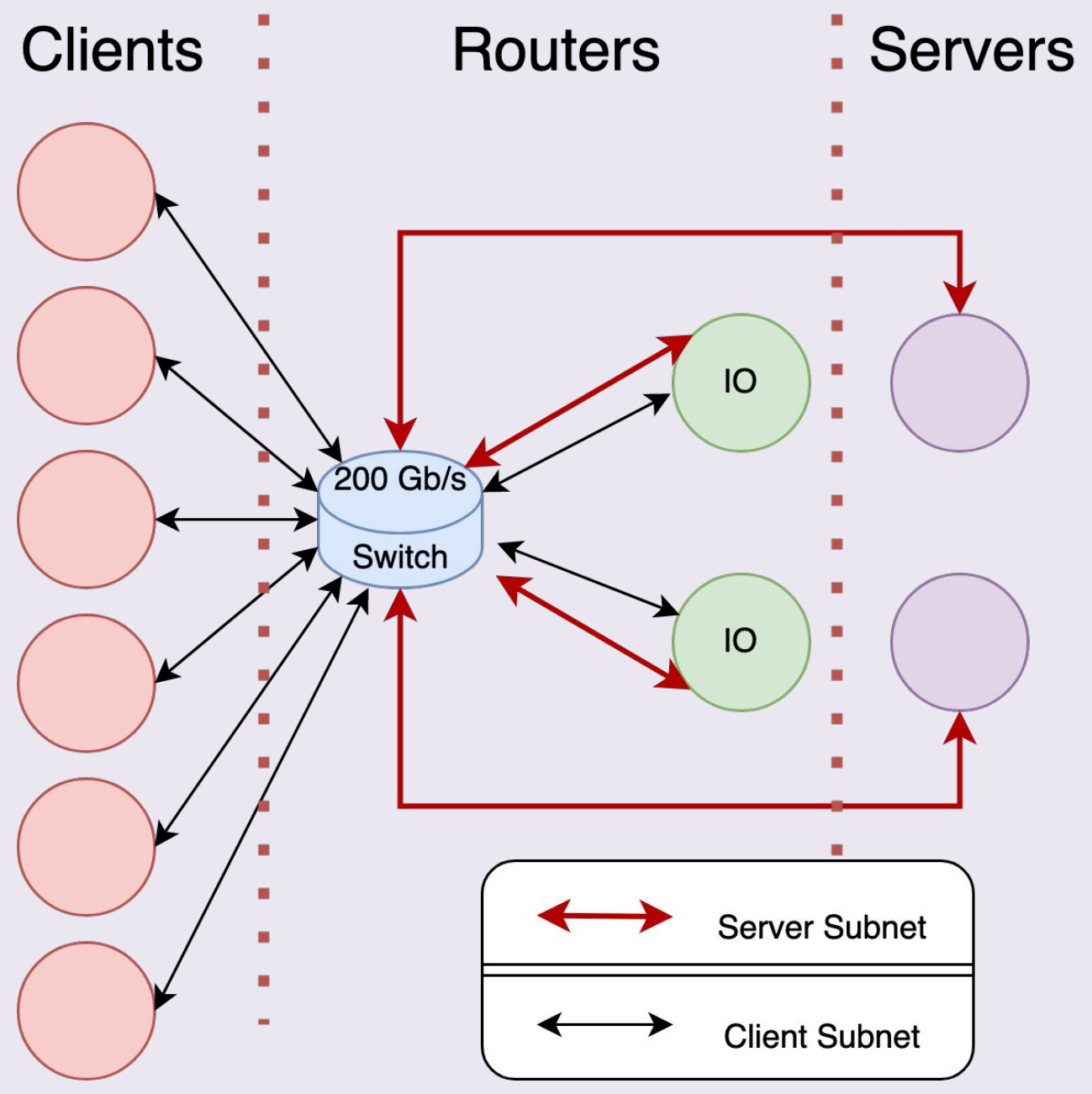

Figure 1. Routed

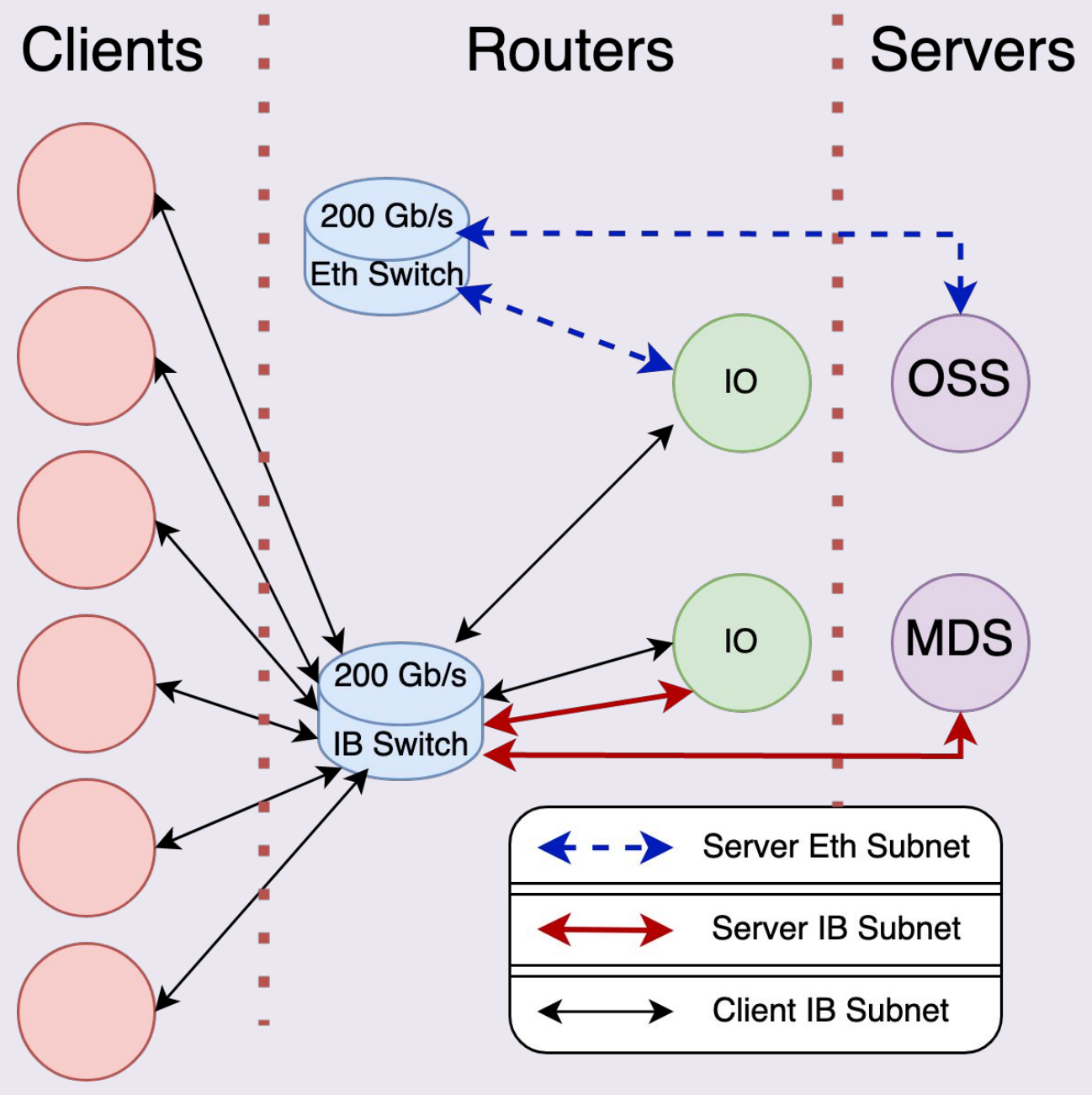

Figure 2. Split Routed

Topology Evolution

We start with flat InfiniBand and Ethernet topologies. Then we switch to routed Ethernet and InfiniBand networks. Shown above in Figure 2 is the final, split routed Ethernet and InfiniBand topology.

Lustre Network Performance

Lustre Network benchmarking is done with LST to measure bandwidth over the network with different operations and topologies. We run tests with read and write, incrementing message sizes, as well as all-all, all-1, and 1-1 distributions. All topologies are used as described in *Network Topologies* above. Shown below is the average bandwidth of all clients connecting to one server at the same time with a 1MB message size.

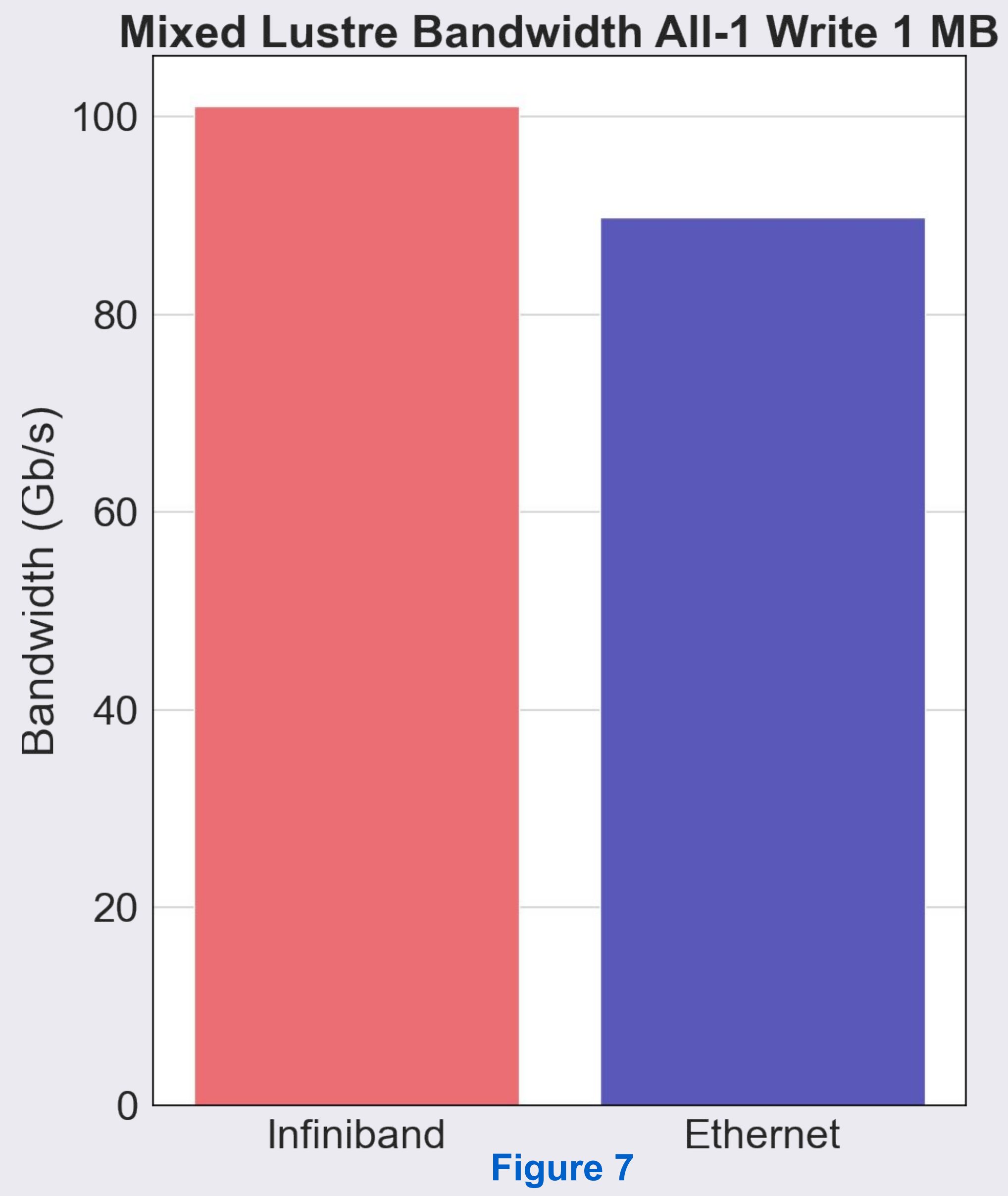

Figure 7

RDMA Performance

We run `perftest` benchmarks on both flat Ethernet and flat InfiniBand topologies. The tests perform send, read, and write operations over IPv4 and IPv6 for RDMA to measure bandwidth and latency on the cluster with a 1-1 distribution. With Ethernet, RoCE is used to perform RDMA operations.

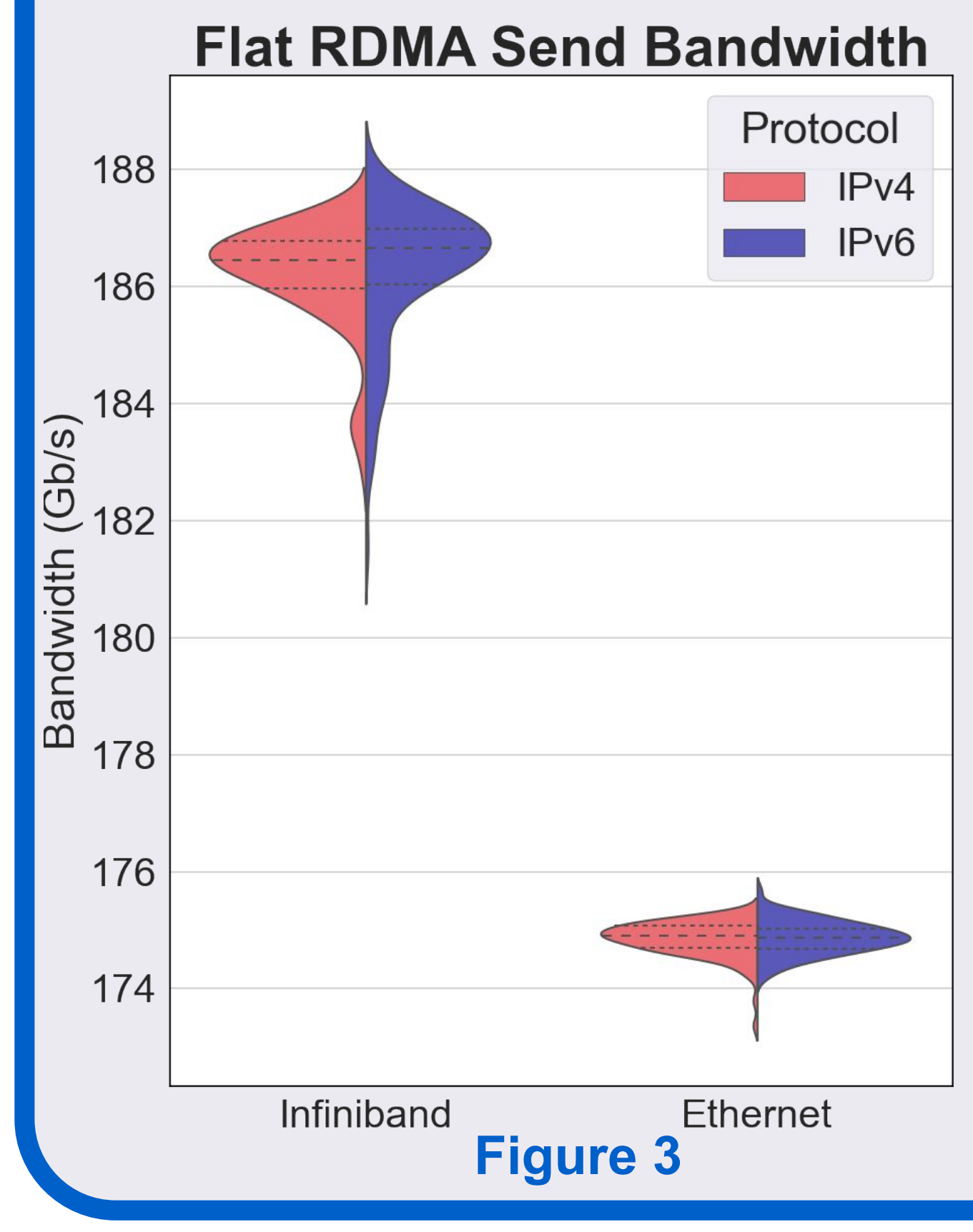

Figure 3

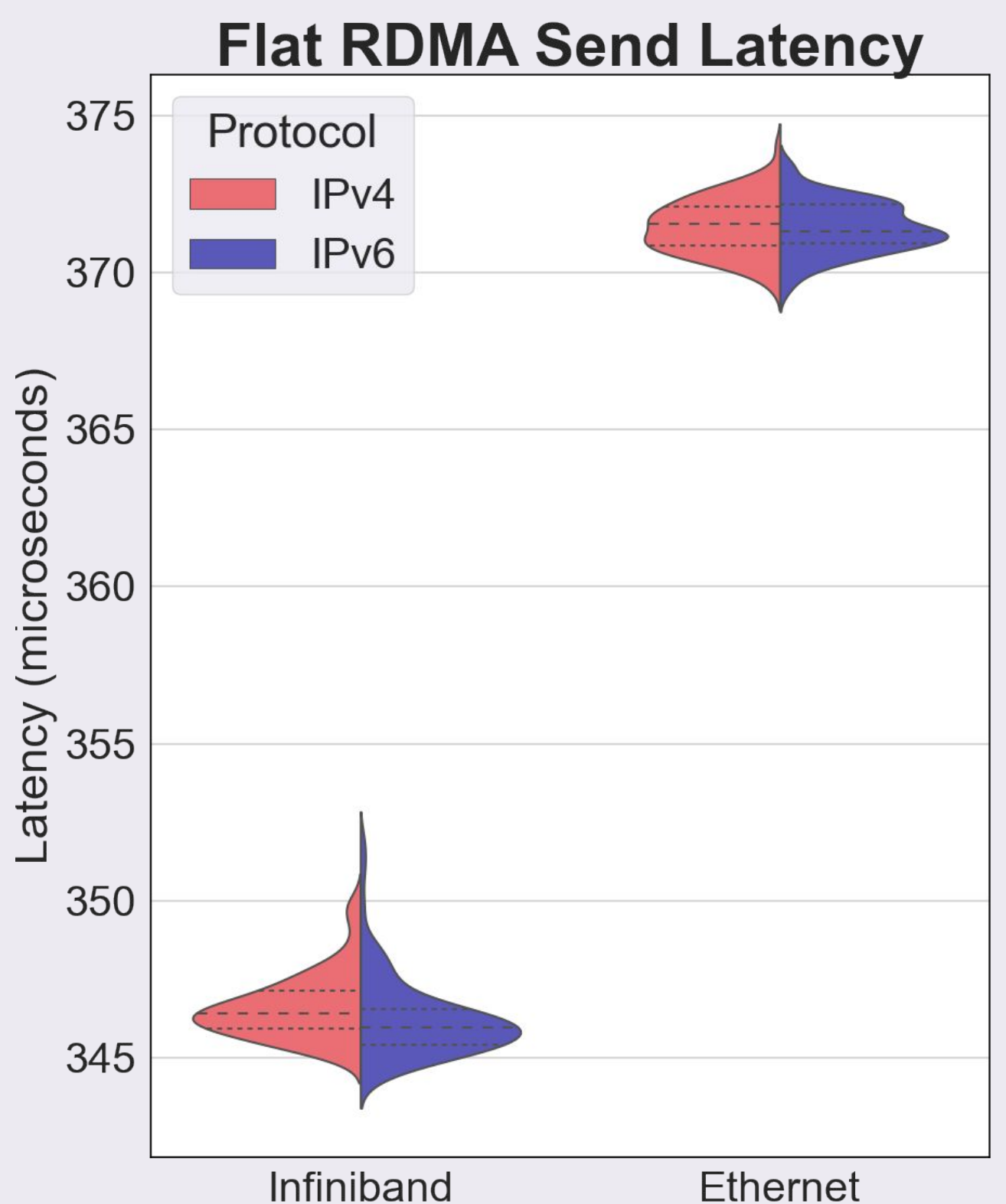

Figure 4

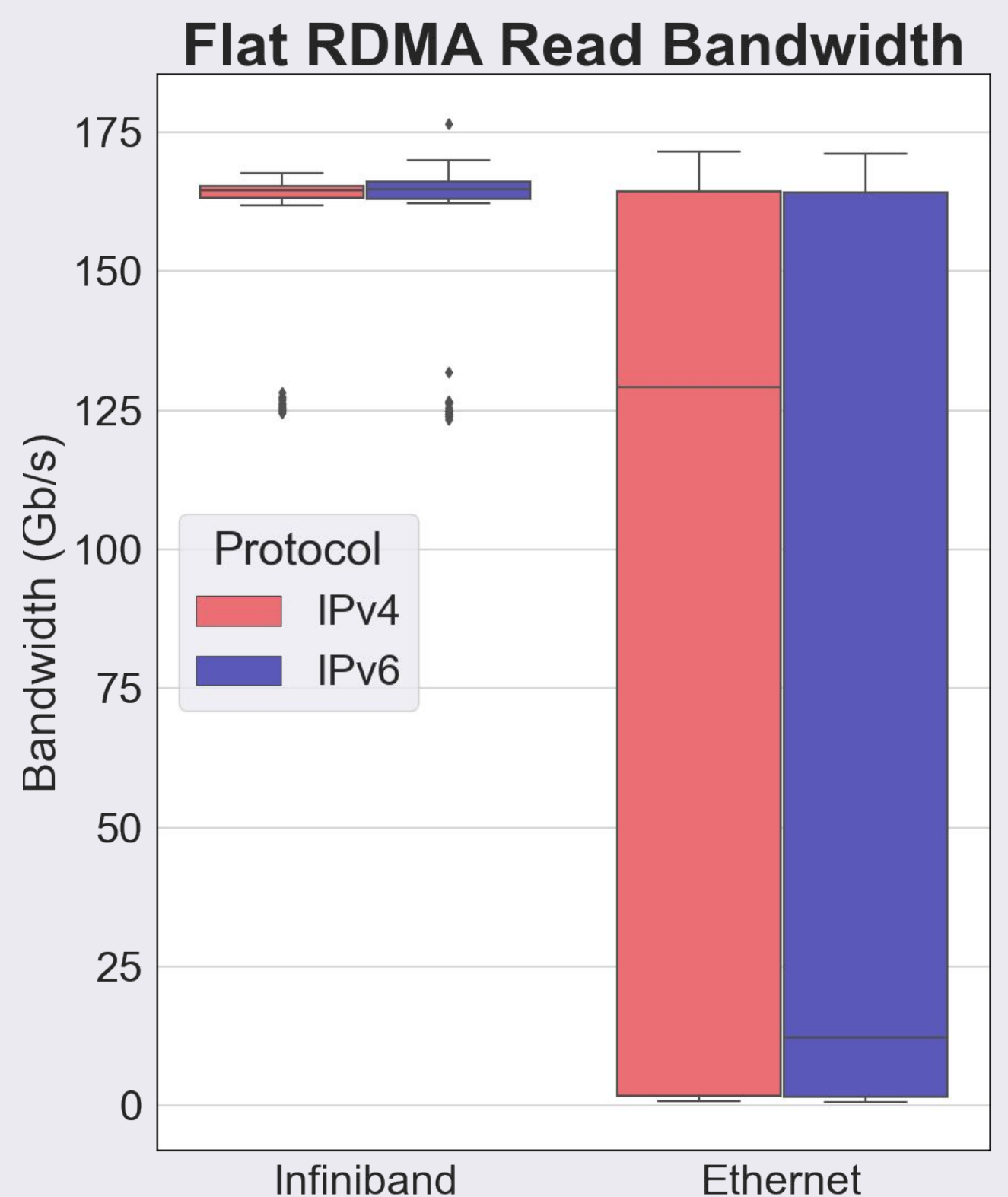

Figure 5

TCP Performance

We use `iperf` to benchmark TCP bandwidth. This test performs send operations with IPv4 and IPv6 in a 1-1 distribution pattern. We run this test for all topologies as described in *Network Topologies*. When using InfiniBand we use IPoIB to perform TCP operations.

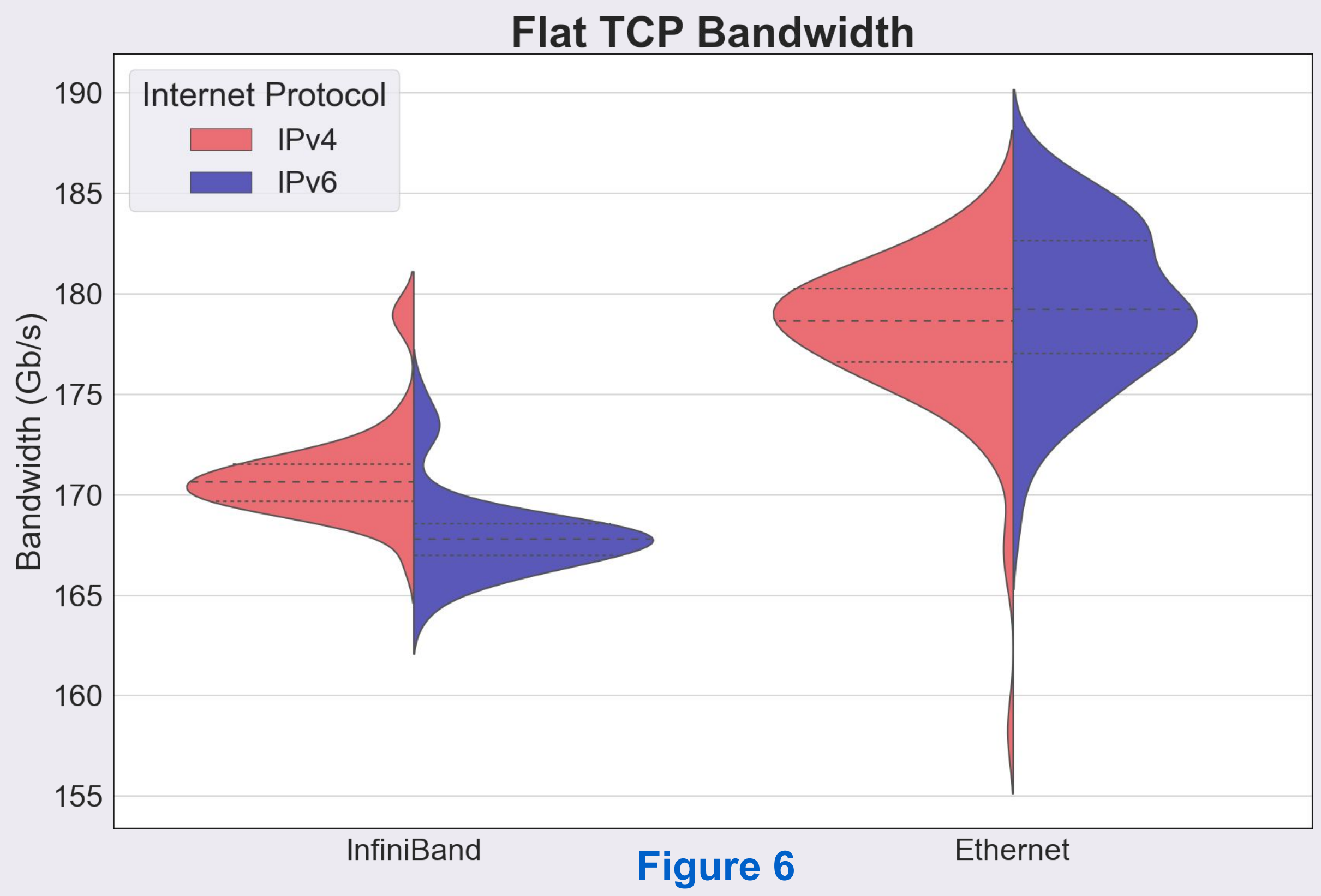

Figure 6

Conclusions

SUMMARY

- The RDMA bandwidth of RoCE is close but not equivalent to current RDMA InfiniBand performance
- With further tuning RoCE could come close to matching InfiniBand
- InfiniBand is slightly more stable and performant than RoCE
- RoCE requires more tuning and configuration than InfiniBand
- Ethernet has better TCP/IP bandwidth than InfiniBand

FUTURE RESEARCH

- Lustre with IPv6 (when supported)
- Ultra Ethernet
- Further RoCE/Lustre tuning
- MPI testing
- Benchmark full Lustre setup with functional MDS and OSS

Managed by Triad National Security, LLC, for the U.S. Department of Energy's NNSA.

LA-UR-23-28982